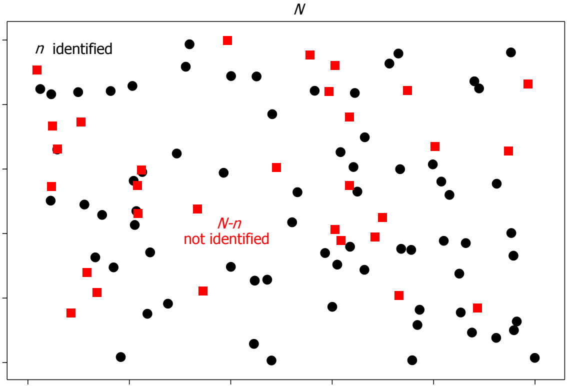
The Zelterman Estimate of Population Size under Heterogeneity

Dankmar Böhning

Quantitative Biology and Applied Statistics, School of Biological Sciences
University of Reading

`IBC, Dublin, 13-18 July 2008`

a population has N units of which n are identified by some mechanism (trap, register, police database, ...)



Formulation of the Problem

- probability of identifying an unit is $(1 - p_0)$
- so that $N = (1 - p_0)N + p_0 N = n + p_0 N$
- and the **Horvitz-Thompson** estimator follows:

$$\hat{N} = \frac{n}{1 - p_0}$$

- usually an estimate of p_0 is required

Formulation of the Problem as Frequencies of Frequencies

a common setting for estimating p_0 is the **Frequencies of Frequencies** setting:

- the identifying mechanism provides a count Y of **repeated identifications** (w.r.t. to a reference period)
- leading to frequencies $f_1, f_2, ..., f_m$ of the counts $1, 2, .., m$ (m is the largest observed count)
- zero counts are **not** observed: hence f_0 **is unknown**
- Recall that $N = f_0 + n = f_0 + f_1 + f_2 + ... + f_m$, so that $\hat{f}_0$ leads to $\hat{N}$

Application Areas

- Epidemiology and Medicine
- Biology and Agriculture
- Social Science and Criminology
- Research on Terrorism

Hser's Data on Estimating Hidden Intravenous Drug Users in Los Angeles 1989

- intravenous drug users in L.A. county were entered into the California Drug Abuse Data System (CAL-DADS)
- the data below refer to the frequency distribution of the episode count per drug user in 1989

the frequency distribution of the **episode count per drug user** for the year 1989:

f_0	f_1	f_2	f_3	f_4	f_5	f_6
-	11,982	3,893	1,959	1,002	575	340

f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	n
214	90	72	36	21	14	20,198

Del Rio Vilas's Data on Estimating Hidden Scrapie in Great Britain 2005

- sheep is kept in holdings in Great Britain (and elsewhere)
- the occurrence of scrapie is monitored in the Compulsory Scrapie Flocks Scheme (CSFS) summarizing abbatoir survey, stock survey and the statutory reporting of clinical cases
- CSFS established since 2004

the frequency distribution of the **scrapie count within each holding** for the year 2005:

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	n
-	84	15	7	5	2	1	2	2	118

Formulation of the Problem and the Idea for its Solution

Suppose we can find some model for the count probabilities

$$p_j = p_j(\lambda)$$

then estimate λ by some method (truncated likelihood) and then use the model for p_0:

$$\hat{N} = \frac{n}{1 - p_0(\hat{\lambda})}$$

Formulation of the Problem and the Idea for its Solution

Only to illustrate: Poisson model for the count probabilities

$$p_j = p_j(\lambda) = \exp(-\lambda)\lambda^j/j!$$

then estimate λ and arrive at:

$$\hat{N} = \frac{n}{1-\hat{p}_0} = \frac{n}{1-\exp(-\hat{\lambda})}$$

Formulation of the Problem and the Idea for its Solution

However: using a simple Poisson model for the count probabilities

$$p_j = p_j(\lambda) = \exp(-\lambda)\lambda^j/j!$$

is **not** appropriate, since

- every unit is different
- there is population heterogeneity

so that more **realistic**

$$p_j = p_j(\lambda) = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

where $\lambda(t)$ stands for the heterogeneity distribution of the Poisson parameter instead of providing an estimate $\hat{\lambda}(t)$ by means of **parametric or nonparametric mixture models** interest is on **two alternatives**:

1. lower bound approach by **Chao** (1987, 1989, *Biometrics*)

$$p_j = \int_0^\infty \exp(-t)t^j/j!\lambda(t)dt$$

   with unknown $\lambda(t)$ for $t > 0$. Then, by the Cauchy-Schwartz inequality:

$$p_1^2 \leq p_0 2p_2 \Leftrightarrow \frac{p_1^2}{2p_2} \leq p_0$$

   leads to **Chao's lower bound estimate**

$$\hat{f}_0 = \frac{f_1^2}{2f_2}$$

2. robust approach of **Zelterman** (1988, *JSPI*)

The Idea of Zelterman (1988)

- he noted that

$$\lambda = \frac{\lambda^{j+1}}{\lambda^j} = (j+1)\frac{\lambda^{j+1}/(j+1)!}{\lambda^j/j!}$$

$$\lambda = (j+1)\frac{Po(j+1;\lambda)}{Po(j;\lambda)}$$

- leading to the proposal

$$\hat{\lambda}_j = (j+1)\frac{f_{j+1}}{f_j}$$

- and in particular for $j = 1$

$$\hat{\lambda} = \hat{\lambda}_1 = 2\frac{f_2}{f_1}$$

$\hat{\lambda} = 2\frac{f_2}{f_1}$ is **robust** in the sense that

- it is **not affected** by any changes in counts larger than 2
- count distribution need only to behave **like** a Poisson for counts of 1 or 2

Introduction

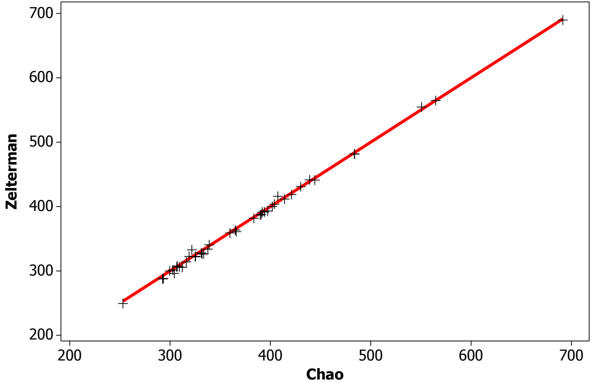
Some Applications

Solutions to the Population Size Problem

Some Recent Results on Zelterman Estimation

- How are Chao's and Zelterman's Estimator related?
- Zelterman as MLE
- Zelterman can be extended to case data
- Zelterman extended to higher counts
- Some simulation results

Zelterman larger than Chao? Carothers capture-recapture data on the number of taxis in Edinburgh (42 sampling occasions)



Zelterman larger than Chao?

- $$\hat{N}_Z = \frac{n}{1-\exp(-\hat{\lambda})} = n + \frac{n}{\exp(\hat{\lambda})-1} \approx n + \frac{n}{1+\hat{\lambda}+\frac{1}{2}\hat{\lambda}^2-1}$$

- $$= n + \frac{n}{\hat{\lambda}+\frac{1}{2}\hat{\lambda}^2} = n + \frac{n}{\frac{2f_2}{f_1}+\frac{1}{2}\left(\frac{2f_2}{f_1}\right)^2} = n + \left(\frac{f_1^2}{2f_2}\right)\frac{n}{f_1+f_2}$$

- $$\geq n + \left(\frac{f_1^2}{2f_2}\right) = \hat{N}_C$$

- **yes,** if $\hat{\lambda}$ is **small** (Böhning and Brittain *SJOS* 2008)

Zelterman, Chao and simple Poisson MLE for the four data sets

		$\hat{N}$				
Example	n	MLE	Chao	Zelterman	$\frac{f_2}{f_1}$	$\frac{n}{f_1+f_2}$
Scrapie	118	188	353	393	0.18	1.19
Drug Use L.A.	20,198	26,425	38,637	42,268	0.33	1.27

MLE:

$$\hat{N} = \frac{n}{1 - p_0(\hat{\lambda})}$$

where $\hat{\lambda}$ is MLE under homogenuous Poisson

Zelterman Estimation offers Flexibility

Zelterman estimate truncates all counts different from 1 or 2: write

$$1 - p = p_1 = \frac{\exp(-\lambda)\lambda}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2} = \frac{1}{1 + \lambda/2}$$

$$p = p_2 = \frac{\exp(-\lambda)\lambda^2/2}{\exp(-\lambda)\lambda + \exp(-\lambda)\lambda^2/2} = \frac{\lambda/2}{1 + \lambda/2}$$

and consider associated **binomial** log-likelihood

$$f_1 \log(p_1) + f_2 \log(p_2) = f_1 \log(1 - p) + f_2 \log(p)$$

which is maximized for $\hat{p} = \hat{p}_2 = \frac{f_2}{f_1+f_2}$, or

$$\hat{\lambda} = \frac{2\hat{p}_2}{1 - \hat{p}_2} = \frac{2f_2}{f_1}$$

Zelterman Estimation offers Flexibility

a likelihood framework offers generalizations:

- (correct) variance estimate of the Zelterman estimator (Fisher information) (Böhning 2008, *Statistical Methodology*)
- extension of the estimator for **case data**
- incorporation of **covariates** (binomial logistic regression with log-link function to the Poisson parameter) (Böhning and van der Heijden 2008 *Ann. Appl. Statist.*)
- efficiency

Zelterman Estimation: Extension to Case Data

Table: *Illustration of Case Data with Individual Recapture Counts*

Unit i	Count y_i	δ_i	Sex_i	Age_i
1	1	0	Male	34
2	2	1	Male	21
3	1	0	Female	34
4	3	-	Male	19
5	2	1	Female	17
6	1	0	Female	26
...	...	...	...	...

Zelterman Estimation offers Flexibility

Binomial likelihood for **grouped** data

$$f_1 \log(1-p) + f_2 \log(p)$$

becomes for **case** data

$$\sum_i (1-\delta_i) \log(1-p) + \delta_i \log(p)$$

which becomes with covariate information on case i

$$p_i = \frac{\exp(\beta^T \mathbf{x}_i)}{1+\exp(\beta^T \mathbf{x}_i)}$$

a **logistic regression** model

Zelterman Estimation offers Flexibility

covariate information on case i

$$p_i = \frac{\exp(\beta^T \mathbf{x}_i)}{1 + \exp(\beta^T \mathbf{x}_i)}$$

compare with parameterization in capture probability λ

$$p_i = \frac{\lambda_i/2}{1 + \lambda_i/2}$$

it follows that

$$\lambda_i = 2\exp(\beta^T \mathbf{x}_i)$$

and the **generalization of the Horvitz-Thompson estimator** is

$$\sum_{i=1}^{n} \frac{1}{1 - \exp(-2e^{\beta^T \mathbf{x}_i})}$$

Generalizing the Idea of Zelterman: Improving Efficiency

- not only

$$\lambda = 2\frac{Po(2;\lambda)}{Po(1;\lambda)}$$

- but also

$$\lambda = \lambda \overbrace{\left(\frac{\lambda + \lambda^2/2!}{\lambda + \lambda^2/2!}\right)}^{1} = \frac{2\lambda^2/2! + 3\lambda^3/3!}{\lambda + \lambda^2/2!}$$

$$= \frac{2Po(2;\lambda) + 3Po(3;\lambda)}{Po(1;\lambda) + Po(2;\lambda)}$$

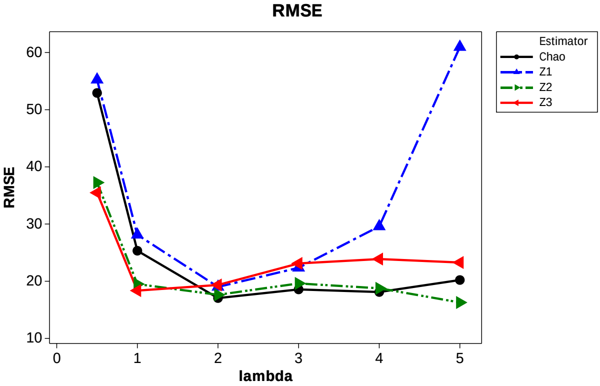
- $$\lambda = 2\frac{Po(2;\lambda)}{Po(1;\lambda)} = \frac{2Po(2;\lambda)+3Po(3;\lambda)}{Po(1;\lambda)+Po(2;\lambda)}$$
  $$= \frac{2Po(2;\lambda)+3Po(3;\lambda)+4Po(4;\lambda)}{Po(1;\lambda)+Po(2;\lambda)+Po(3;\lambda)} = \ldots$$
- leads to the **proposal**

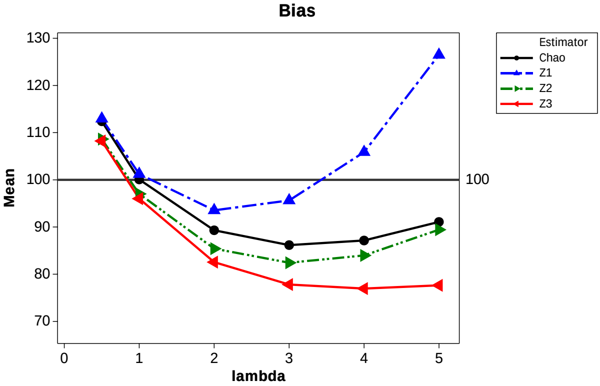
$$\hat{\lambda} = \hat{\lambda}_1 = 2\frac{f_2}{f_1}, \hat{\lambda}_2 = \frac{2f_2+3f_3}{f_1+f_2}, \hat{\lambda}_3 = \frac{2f_2+3f_3+4f_4}{f_1+f_2+f_3}, \ldots$$

- $$\hat{N}_i = \frac{n}{1-\exp(-\hat{\lambda}_i)}$$

Simulation Experiment

- **Goal**: Compare $\hat{N}_j = \frac{n}{1-\exp(-\hat{\lambda}_j)}$ and Chao's estimator $\hat{N}_C = n + \frac{f_1^2}{2f_2}$
- **count data:** f_j arise from $0.5Po(j; 0.5) + 0.5Po(j; \lambda)$ for $\lambda = 1, 2, ..., 7$ and $j = 0, 1, 2, ...$
- **population size:** $N = f_0 + f_1 + ... = 100$
- f_0 is **truncated**
- N estimated using $\hat{N}_1$, $\hat{N}_2$, $\hat{N}_3$ and $\hat{N}_C$


RMSE
Estimator
Chao
Z1
Z2
Z3
RMSE
60
50
40
30
20
10
0
1
2
3
4
5
lambda


Bias
Mean
lambda
130
120
110
100
90
80
70
0
1
2
3
4
5
100
Estimator
Chao
Z1
Z2
Z3

Main Conclusions

- Estimators of Chao and Zelterman closely related
- however: Zelterman Estimation offers more flexibility because ...
  - increasing its efficiency via truncated likelihood
  - incorporation of case data
  - incorporation of prior information by means of covariates

Generalizing the idea of Zelterman: truncation or censoring?

- disadvantage of conventional Zelterman: uses only f_1 and f_2
- idea of truncation: ignore all counts different from 1 and 2
- idea of censoring: use marginal likelihood for all counts of 2 and larger
- $$p_1 = P(Y = 1) = \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \lambda$$
- $$p_{2+} = P(Y > 1) = \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} [\lambda^2/2! + \lambda^3/3! + \dots]$$
$$= 1 - \frac{\exp(-\lambda)}{1 - \exp(-\lambda)} \lambda$$

Generalizing the idea of Zelterman: truncation or censoring?

leads to the **binomial likelihood**

$$f_1 \log(p_1) + f_{2+} \log(p_{2+})$$

since

$$\hat{p}_1 = f_1/n$$

we have

$$f_1/n = \frac{\exp(-\lambda)}{1-\exp(-\lambda)}\lambda = \frac{1}{\exp(\lambda)-1}\lambda$$

$$\approx \frac{1}{\lambda + \lambda^2/2}\lambda$$

leads to

$$\hat{\lambda}_C = \frac{2(n-f_1)}{f_1}$$

frequently: evidence for a 2-component mixture model

Table: Amount of heterogeneity occurring in the data sets

Example	Non-parametric mixture model
McKendrick	homogeneity
Matthews	**2**-component
Scrapie	**2**-component
Drug Use L.A.	**3**-component
terrorist activity	**6**-component